The Annals of Applied Statistics
2012, Vol. 6, No. 4, 1664–1688
DOI: 10.1214/12-AOAS561

MULTIPLE HYPOTHESIS TESTING ADJUSTED FOR LATENT VARIABLES, WITH AN APPLICATION TO THE AGEMAP GENE EXPRESSION DATA

BY YUNTING SUN[1,2], NANCY R. ZHANG[2] AND ART B. OWEN[1]

Stanford University

In high throughput settings we inspect a great many candidate variables (e.g., genes) searching for associations with a primary variable (e.g., a phenotype). High throughput hypothesis testing can be made difficult by the presence of systemic effects and other latent variables. It is well known that those variables alter the level of tests and induce correlations between tests. They also change the relative ordering of significance levels among hypotheses. Poor rankings lead to wasteful and ineffective follow-up studies. The problem becomes acute for latent variables that are correlated with the primary variable. We propose a two-stage analysis to counter the effects of latent variables on the ranking of hypotheses. Our method, called LEAPP, statistically isolates the latent variables from the primary one. In simulations, it gives better ordering of hypotheses than competing methods such as SVA and EIGENSTRAT. For an illustration, we turn to data from the AGEMAP study relating gene expression to age for 16 tissues in the mouse. LEAPP generates rankings with greater consistency across tissues than the rankings attained by the other methods.

1. Introduction. There has been considerable progress in multiple testing methods for high throughput applications. A common example, coming from biology, is testing which of N genes' expression levels correlate significantly with a scalar variable, which we'll call the primary variable. The primary variable may be an experimentally applied treatment or it may be a covariate such as a phenotype. We will use the gene expression example for concreteness, although it is just one of many instances of this problem.

High throughput experiments may involve thousands or even millions of hypotheses. Because N is so large, serious problems of multiplicity arise. For

Received May 2011; revised April 2012.
[1]Supported by NSF Grant DMS-09-06056.
[2]Supported by NSF Grant DMS-09-06394.
Key words and phrases. EIGENSTRAT, empirical null, surrogate variable analysis.

independent tests, methods based on the false discovery rate [Dudoit and van der Laan (2008)] have been very successful. Attention has turned more recently to dependent tests [Efron (2010)].

One prominent cause of dependency among test statistics is the presence of latent variables. For example, in microarray-based experiments, it is well known that samples processed in the same batch are correlated. Batch, technician and other sources of variation in sample preparation can be modeled by latent variables. Another example comes from genetic association studies, where differences in ancestral history among subjects can lead to false or inaccurate associations. Price et al. (2006) used principal components to extract and correct for ancestral history, in effect modeling the genetic background of the subjects as latent variables. A third example comes from copy number data, where local trends along the genome cause false positive copy number calls [Olshen et al. (2004)]. Diskin et al. (2008) conducted experiments showing that these local trends correlate with the percentage of bases that are guanines or cytokines along the genome, and are caused by differences in the quantity and handling of DNA. These laboratory effects are hard to measure, but can be quantified using a latent variable model. In this paper, we consider latent variables that might even be correlated with the primary variable.

When the primary variable is an experimentally applied treatment, then problematic latent variables are those that are partially confounded with the treatment. Randomization reduces the effects of such confounding, but randomization is not always perfectly applied and batch or other effects may be imbalanced with respect to the treatment [Leek et al. (2010)].

These latent variables have some severe consequences. They alter the level of the hypothesis tests and they induce correlations among multiple tests. Another consequence, that we find especially concerning, is that the latent variables may affect the rank ordering among the N p-values. When high throughput methods are used to identify candidates for further follow-up it is important that the highly ranked items contain as many nonnull cases as possible.

Our approach to this problem uses a rotated model in which we separate the latent variables from the primary variable. We do this by creating two data sets, one in which both primary and latent variables are present and one in which the primary variables are absent. We use the latter data set to estimate the latent variables and then substitute their estimates into the former. Since each gene has its own effect size in relation to the primary variable, the former model is supersaturated. We conduct inference under the setting where the parameter vector relating the genes to the primary variable is sparse, as is commonly assumed in multiple testing situations. Each nonnull hypotheses behaves as an additive outlier, and we then apply an outlier detection method from She and Owen (2011) to find them. We call the method LEAPP, for *l*atent *e*ffect *a*djustment after *p*rimary *p*rojection.

Section 2 presents our notation and introduces LEAPP along with several other related models, including SVA [Leek and Storey (2008)] and EIGENSTRAT [Price et al. (2006)], to which we make comparisons. Section 3 shows via simulation that LEAPP generates better rankings of the nonnull hypotheses than one would get by either ignoring the latent variables, by SVA, or by EIGENSTRAT. EIGENSTRAT estimates the latent variables (by principal components) without first adjusting for the primary variable. LEAPP outperforms it when the latent variable is weaker than the primary. EIGENSTRAT does well in simulations with weak primary variables, which matches the setting that motivated it. Still it is interesting to learn that it does not extend well to problems with strong primary variables. SVA estimates the primary variable's coefficients without first adjusting for correlation between the primary and latent variables. LEAPP outperforms it when the latent and primary variables are correlated.

Section 4 compares the methods on the AGEMAP data of Zahn et al. (2007). The primary variable there is age. While we do not know the truly nonnull genes for this problem, we have a proxy. The data set has 16 subsets, each from a different tissue type. We find that LEAPP gives gene lists with much greater overlap among tissues than the gene lists achieved by the other methods. Our conclusions are in Section 5. We include some brief remarks on calibration of the p-values themselves as opposed to the rank ordering which is the primary focus of this paper. Some theory is given in the Appendix for a simplified version of LEAPP. The specific rotation matrix used does not affect our answer. For the case of one latent variable and no covariates, the simplified LEAPP consistently estimates the latent structure. We also get a bound for the sum of squared coefficient errors when the effects are sparse.

2. Notation and models. In this section we describe the data model and introduce the parameters and latent variables that arise. Then we describe our LEAPP proposal which is based on a series of reductions from a heteroscedastic multivariate regression including latent factors to a single linear regression problem with additive outliers and known error variance. We also describe EIGENSTRAT and SVA, to which we make comparisons, and then survey several other published methods for this problem.

2.1. *Data, parameters, latent variables and tests.* The data we observe are a response matrix $Y \in \mathbb{R}^{N\times n}$ and a variable of interest $g \in \mathbb{R}^n$, which we call the primary variable. In an expression problem Y_{ij} is the expression level of gene i for subject j. Very often the primary variable g is a group variable taking just two values, such as ± 1 for a binary phenotype, then linearly transformed to have mean 0 and norm 1. The quantity g_j can also be a more general scalar, such as the age of subject j.

We are interested to know which genes, if any, are linearly associated with the variable g. We capture this linear association through the $N \times n$ matrix

γg^T, where γ is a vector of N coefficients. When most genes are not related to g, then γ is sparse.

Often there are covariates X other than g that we should adjust for. The covariate term is βX^T where β contains coefficients. The latent variables that cause tests to be mutually correlated are assumed to take an outer product form UV^T. Neither U nor V is observed. Finally, there is observational noise with a variance that is allowed to be different for each gene, but assumed to be constant over subjects.

The full data model is

$$Y = \gamma g^\mathsf{T} + \beta X^\mathsf{T} + UV^\mathsf{T} + \Sigma E \tag{2.1}$$

for variables

$$\begin{aligned}
Y &\in \mathbb{R}^{N\times n} && \text{response values,}\\
g &\in \mathbb{R}^{n\times 1} && \text{primary predictor, that is, treatment, with } g^\mathsf{T}g = 1,\\
\gamma &\in \mathbb{R}^{N\times 1} && \text{primary parameter, possibly sparse,}\\
X &\in \mathbb{R}^{n\times s} && s \text{ covariates (e.g., sex) per subject,}\\
\beta &\in \mathbb{R}^{N\times s} && s \text{ coefficients, including per gene intercepts,}\\
U &\in \mathbb{R}^{N\times k} && \text{latent, nonrandom rows (e.g., genes),}\\
V &\in \mathbb{R}^{n\times k} && \text{latent, independent rows (e.g., subjects),}\\
E &\sim \mathcal{N}(0, I_N \otimes I_n) && \text{noise}
\end{aligned}$$

and

$$\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_N) \qquad \text{standard deviations}$$

with dimensions

$$\begin{aligned}
n &&& \text{number of arrays/subjects,}\\
N &\gg n && \text{number of genes,}\\
s &\ll n && \text{number of covariates}
\end{aligned}$$

and

$$k \geq 1 \qquad \text{latent dimension.}$$

After adjusting for X, the genes are correlated through the action of the latent portion UV^T of the model. They may have unequal variances, through both Σ and U. We adopt the normalization $\mathbb{E}(V^\mathsf{T}V) = I_k$. It is possible to generalize the model to have a primary variable g of dimension larger than one, but we focus on the case of a single primary variable.

We pay special attention to the case of $k = 1$ latent variable. The algorithm is the same for all values of k. But, when $k = 1$, the dependence between the variable g of interest and the latent variable V can be summarized by a single correlation coefficient $\rho = g^\mathsf{T}V/\sqrt{V^\mathsf{T}V}$ which aids interpretation.

Writing (2.1) in terms of indices yields

$$Y_{ij} = \gamma_i g_j + \beta_i^\mathsf{T} X_j + U_i^\mathsf{T} V_j + \sigma_i \varepsilon_{ij}, \qquad 1 \le i \le N, 1 \le j \le n. \tag{2.2}$$

Here β_i and U_i are the ith rows of β and U, respectively, as column vectors. Similarly, X_j and V_j are the jth rows of X and V, σ_i is the ith diagonal element of Σ and ε_{ij} is the ij element of E.

Our LEAPP proposal is based on a series of reductions described next. In outline, we first split the data into two parts, one of which is completely free of the primary variable. We then estimate some properties of the latent variable model from that primary-free data. Finally, we use those estimated quantities in the part of the data which does contain the primary variable to identify genes related to the primary variable.

2.2. *Data rotation.* We begin by choosing an orthogonal matrix $O \in \mathbb{R}^{n\times n}$ such that $g^\mathsf{T} O^\mathsf{T} = (\eta, 0, 0, \ldots, 0) \in \mathbb{R}^{1\times n}$ where $\eta = \|g\| > 0$. Without loss of generality, we assume that the primary predictor has been scaled so that $\eta = 1$. A convenient choice for O is the Householder matrix $O = I_n - 2\kappa\kappa^\mathsf{T}$, where $\kappa = (g - e_1)/\|g - e_1\|_2$ and $e_1 = (1, 0, \ldots, 0)^\mathsf{T}$.

Using O, we construct the *rotated* model

$$Y^{(r)} \equiv YO^\mathsf{T} = \gamma g^\mathsf{T} O^\mathsf{T} + \beta X^\mathsf{T} O^\mathsf{T} + UV^\mathsf{T} O^\mathsf{T} + \Sigma E O^\mathsf{T} \tag{2.3}$$

$$\equiv \gamma g^{(r)\mathsf{T}} + \beta X^{(r)\mathsf{T}} + UV^{(r)\mathsf{T}} + \Sigma E^{(r)}, \tag{2.4}$$

where $g^{(r)}$, $X^{(r)}$, $V^{(r)}$ and $E^{(r)}$ are rotated versions of g, X, V and E, respectively. For each major transformation of the data, a new mnemonic superscript will be introduced. Some superscripts use the same letter also used as a data dimension, but the usages are distinct enough that one will not be mistaken for the other.

Notice that $E^{(r)} = EO^\mathsf{T} \overset{d}{=} E$, because $E \sim \mathcal{N}(0, I_N \otimes I_n)$. By construction, $g^{(r)} = (1, 0, \ldots, 0)$. Therefore, the model for $Y_{ij}^{(r)}$ is different depending on whether $j = 1$ or $j \ne 1$:

$$Y_{i1}^{(r)} = \beta_i^\mathsf{T} X_1^{(r)} + U_i^\mathsf{T} V_1^{(r)} + \gamma_i + \sigma_i \varepsilon_{i1}^{(r)} \tag{2.5}$$

and

$$Y_{ij}^{(r)} = \beta_i^\mathsf{T} X_j^{(r)} + U_i^\mathsf{T} V_j^{(r)} + \sigma_i \varepsilon_{ij}^{(r)}, \qquad j = 2, \ldots, n, \tag{2.6}$$

where $\varepsilon_{ij}^{(r)}$ is the (i, j)th element of $E^{(r)}$.

The rotated model concentrates the primary coefficients γ_i in the first column of $Y^{(r)}$. Our approach is to base tests and estimates of γ_i on equation (2.5). We need to substitute estimates for unknown quantities σ_i, β_i and U_i in (2.5). The estimates come from the model in equation (2.6).

This rotated approach has some practical advantages: First, we do not need to iterate between applying equations (2.5) and (2.6). Instead we use

(2.6) once to estimate unknowns U, σ and β and then use (2.5) once to judge γ_i. Second, the last $n-1$ columns of $Y^{(r)}$, and hence estimates $\widehat{\sigma}$, $\widehat{\beta}$, and $\widehat{U}$, are statistically independent of the first column. Third, problems (2.5) and (2.6) closely match settings for which there are usable methods as described next.

Using estimates $\widehat{\sigma}_i$, $\widehat{U}_i$ and $\widehat{\beta}_i$ from (2.6) described below, we may write (2.5) as

$$Y_{i1}^{(r)} - \widehat{\beta}_i^{\mathsf{T}} X_1^{(r)} = \widehat{U}_i^{\mathsf{T}} V_1^{(r)} + \gamma_i + \widehat{\sigma}_i \varepsilon_{i1}^{(r)}. \tag{2.7}$$

The right-hand side of equation (2.7) is a regression with measurement errors in the predictors $\widehat{U}_i$, mean-shift outliers γ_i and unequal error variances. We will use the Θ–IPOD algorithm of She and Owen (2011), adjusted to handle unequal σ_i, to get our estimate of γ_i.

Before describing Θ–IPOD we show how to get the estimates $\widehat{\beta}_i$, $\widehat{U}_i$ and $\widehat{\sigma}_i$ from the criss-cross regression algorithm of Gabriel and Zamir (1979). Criss-cross regression will also produce an estimate of $V_j^{(r)}$ for $j \geq 2$, but those vectors do not play a role in (2.7).

2.3. *Estimating U, β and Σ.* We get our estimates of U_i, β_i and σ_i from the last $n-1$ columns of the data set. Let $Y^{(\ell)}$, $X^{(\ell)}$, $V^{(\ell)}$ and $E^{(\ell)}$ be the *last* $n-1$ columns of $Y^{(r)}$, $X^{(r)}$, $V^{(r)}$ and $E^{(r)}$, respectively. Then the model for the last $n-1$ columns of the data is

$$Y^{(\ell)} = \beta X^{(\ell)\mathsf{T}} + U V^{(\ell)\mathsf{T}} + \Sigma E^{(\ell)}. \tag{2.8}$$

Notice that the quantities β, U and Σ in (2.8) are the same as those in the original model (2.1) because the steps taken so far operate on columns of Y. We can write $Y^{(\ell)} = Y^{(r)} D_n$ where $D_n = \binom{0}{I_{n-1}} \in \mathbb{R}^{n\times(n-1)}$ and similarly for $X^{(\ell)}$ and $V^{(\ell)}$. The matrix D_n deletes the first column out of n in the matrix that it follows.

We adopt an iterative approach based on (2.8) that alternates between updating $\widehat{\Sigma}$ and updating the quantities $\widehat{\beta}$, $\widehat{U}$ and $\widehat{V}^{(\ell)}$ given $\widehat{\Sigma}$. The update for $\widehat{\Sigma}$ is

$$\widehat{\Sigma} = \left(\frac{1}{n-1}\operatorname{diag}(\widehat{\varepsilon}\widehat{\varepsilon}^{\mathsf{T}})\right)^{1/2} \qquad \text{where } \widehat{\varepsilon} = Y^{(\ell)} - \widehat{\beta} X^{(\ell)} - \widehat{U}\widehat{V}^{(\ell)\mathsf{T}}. \tag{2.9}$$

That is, $\widehat{\sigma}_i^2$ is simply the mean squared error of a regression for the ith gene.

Given $\widehat{\Sigma}$, we *standardize* the last $n-1$ columns, yielding $Y^{(s\ell)} = \widehat{\Sigma}^{-1} Y^{(\ell)}$. In terms of the other variables,

$$Y^{(s\ell)} = \beta^{(s)} X^{(\ell)\mathsf{T}} + U^{(s)} V^{(\ell)\mathsf{T}} + E^{(s\ell)}, \tag{2.10}$$

where $\beta^{(s)} = \widehat{\Sigma}^{-1}\beta$, $U^{(s)} = \widehat{\Sigma}^{-1}U$ and $E^{(s\ell)} = \widehat{\Sigma}^{-1}E^{(\ell)}$ are standardized versions of β, U and $E^{(\ell)}$, respectively.

Because $\Sigma^{-1}E^{(\ell)}$ has IID Gaussian entries, equation (2.10) closely matches the criss-cross regression model of Gabriel and Zamir (1979). Criss-cross regression for a matrix of data sums three outer products: row based features (with column coefficients), column based features (with row coefficients), and a low rank factor model with latent rows and columns.

We fit a criss-cross regression by first estimating $\beta^{(s)}$ by least squares regression:

$$\widehat{\beta}^{(s)} = Y^{(s\ell)}X^{(\ell)}(X^{(\ell)\mathsf{T}}X^{(\ell)})^{-1}.$$

Then we estimate $U^{(s)}$ and $V^{(\ell)}$ by a truncated singular value decomposition (SVD) of rank k applied to the residuals $\widehat{\varepsilon}^{(s\ell)} = Y^{(s\ell)} - \widehat{\beta}^{(s)}X^{(\ell)\mathsf{T}}$. We absorb the singular values into $\widehat{U}^{(s)}$ but retain the identity $\widehat{V}^{(\ell)\mathsf{T}}\widehat{V}^{(\ell)} = I_k$.

Our use of criss-cross regression has a latent factor model of the form UV^{T} and terms of the form βX^{T} representing column features with row coefficients. The full criss-cross regression model also allows for terms of the form $Z\delta^{\mathsf{T}}$ that combine row features with column coefficients.

To apply the algorithm, we need a starting point for the iteration and a value of k. We start with $\widehat{\Sigma} = I_N$. We have assumed that the rank k for the latent variables is known. When it must be estimated from the data, we follow Leek and Storey (2008) in using the method of Buja and Eyuboglu (1992), as described in Section 2.5.

Criss-cross regression gives us estimates $\widehat{\Sigma}$, $\widehat{\beta}^{(s)}$ and $\widehat{U}^{(s)}$. We can estimate $\widehat{\beta}$ by $\widehat{\Sigma}^{1/2}\widehat{\beta}^{(s)}$ and $\widehat{U}$ by $\widehat{\Sigma}^{1/2}\widehat{U}^{(s)}$. We will use these estimates normalized by $\widehat{\sigma}_i$ and so it is also possible to work with $\widehat{\beta}^{(s)}$ and $\widehat{U}^{(s)}$ themselves.

2.4. *Gene identification.* Now we return to the first column of the rotated data matrix which contains the effects of the primary variable. If we divide $Y_{i1}^{(r)}$ by σ_i, we get

$$\frac{Y_{i1}^{(r)}}{\sigma_i} = \frac{\beta_i^{\mathsf{T}}}{\sigma_i}X_1^{(r)} + \frac{U_i^{\mathsf{T}}}{\sigma_i}V_1^{(r)} + \frac{\gamma_i}{\sigma_i} + \varepsilon_{i1}^{(r)}, \qquad i = 1, \ldots, N. \tag{2.11}$$

For our purposes, equation (2.11) can be cast as a regression of standardized variables on k predictors U_i/σ_i with coefficient vector $V_1^{(r)} \in \mathbb{R}^k$, with additive outliers γ_i/σ_i and offsets $\beta_i^{\mathsf{T}}X_1^{(r)}/\sigma_i$. Though σ_i and β_i and U_i are unknown, we have estimates of them from the previous section.

We use those estimates to construct the *primary* variable regression model

$$Y_i^{(p)} = U_i^{(p)\mathsf{T}}V_1^{(p)} + \gamma_i^{(p)} + \varepsilon_i^{(p)} \tag{2.12}$$

with response $Y_i^{(p)} = (Y_{i1}^{(r)} - \widehat{\beta}_i^{\mathsf{T}}X_1^{(r)})/\widehat{\sigma}_i$, predictors $U_i^{(p)} = \widehat{U}_{i1}^{(r)}/\widehat{\sigma}_i = \widehat{U}_{i1}^{(s)}$, coefficient vector $V_1^{(p)} = V_1^{(r)}$, additive outliers $\gamma_i^{(p)} = \gamma_i/\widehat{\sigma}_i$, and error $\varepsilon_i^{(p)} = \varepsilon_{i1}^{(r)}\sigma_i/\widehat{\sigma}_i$.

The Θ–IPOD algorithm of She and Owen (2011) is designed to estimate a regression coefficient in the presence of additive outliers as well as to identify which observations are outliers. In the present context, the outliers correspond to genes that are associated with the primary variable.

For a complete description of Θ–IPOD see She and Owen (2011), who also cite related work in the robust regression literature. Here we give a brief account of the main points.

The primary variable model (2.12) could be fit by minimizing $\|Y^{(p)} - U^{(p)}V_1^{(p)}\|_2^2 + \lambda\|\gamma^{(p)}\|$ over $V_1^{(p)}$ and $\gamma^{(p)}$. Large enough penalties $\lambda > 0$ would yield a sparse estimate of $\gamma^{(p)}$ which is desirable because the model has $N+k$ parameters and only N observations.

The natural algorithm to minimize the sum of squared errors with an L_1 penalty on the additive outlier coefficients alternates between two steps. One step estimates the additive outlier effects by soft thresholding residuals from a least squares regression. The other step does the least squares regression after first subtracting the estimated outlier effects. She and Owen (2011) found that while soft thresholding is not robust, simply changing the algorithm to do hard thresholding proved to be very robust. Their algorithm also takes account of the leverage values in least squares regression. The algorithm requires a choice for λ. They used a modified BIC statistic from Chen and Chen (2008).

Our statistic for testing $H_{i0}: \gamma_i = 0$ is

$$T_i = \frac{Y_i^{(p)} - U_i^{(p)\mathsf{T}}\widehat{V}_1}{\widehat{\tau}}, \tag{2.13}$$

where $\widehat{V}_1$ is the Θ–IPOD estimate of $V_1^{(p)}$ and $\widehat{\tau}$ is an estimate of the error variance from (2.12). The estimate $\widehat{\tau}$ is the median absolute deviation from the median (MAD) of $Y_i^{(p)} - U_i^{(p)\mathsf{T}}\widehat{V}_1$, with the customary scaling to match the standard deviation for a Gaussian distribution.

For p-values we use $\Pr(|Z| \geq |T_i|)$ where $Z \sim \mathcal{N}(0,1)$. Candidate hypotheses are ranked from most interesting to least interesting by taking the p-values from smallest to largest. This is equivalent to sorting $|T_i|$ from largest to smallest. We consider the quality of this ordering, not whether the p-values are properly calibrated, apart from a brief remark in the conclusions.

The entire LEAPP algorithm is summarized in Figure 1.

We have emphasized the setting in which γ is a sparse vector. When γ is not a sparse vector, then its large components may not be flagged as outliers because the MAD estimate of τ would be inflated due to contamination by γ. In this case, however, we can fall back on a simpler approach to estimating τ. The error $\varepsilon_i^{(p)}$ has variance $\mathbb{E}(\sigma_i^2/\widehat{\sigma}_i^2)$. This variance differs from unity only because of estimation errors in $\widehat{\sigma}_i$. We can then use $\tau^2 = 1$. We can account for fitting s regression parameters to the $n-1$ samples in each

(1) Standardize the primary variable, $g = g/\|g\|$.
(2) Define the rotation matrix $O = I_n - 2\kappa\kappa^\mathsf{T}$ for $\kappa = (g - e_1)/\|g - e_1\|$.
(3) Rotate $Y^{(r)} = YO^\mathsf{T}$ and $X^{(r)} = XO^\mathsf{T}$.
(4) Select the last $n-1$ columns $Y^{(\ell)} = Y^{(r)}D_n$ and $X^{(\ell)} = X^{(r)}D_n$.
(5) Let $\widehat{\beta}^{(s)} = Y^{(\ell)\mathsf{T}}X^{(\ell)}(X^{(\ell)\mathsf{T}}X^{(\ell)})^{-1}$.
(6) Use Buja and Eyuboglu (1992) to estimate the rank k for $Y^{(\ell)} - \widehat{\beta}^{(s)}X^{(\ell)}$.
(7) Set $\widehat{\Sigma} = I_N$.
(8) Iterate to convergence:
  (a) $Y^{(s\ell)} = \widehat{\Sigma}^{-1}Y^{(\ell)}$.
  (b) $\widehat{\beta}^{(s)} = Y^{(s\ell)\mathsf{T}}X^{(\ell)}(X^{(\ell)\mathsf{T}}X^{(\ell)})^{-1}$.
  (c) $\widehat{E}_k^{(s\ell)}$ gets rank k truncated SVD of $\widehat{E}^{(s\ell)} = Y^{(s\ell)} - \widehat{\beta}^{(s)}X^{(\ell)\mathsf{T}}$.
  (d) $\widehat{\Sigma} = (\mathrm{diag}((\widehat{E}^{(s\ell)} - \widehat{E}_k^{(s\ell)})(\widehat{E}^{(s\ell)} - \widehat{E}_k^{(s\ell)})^\mathsf{T})/(n-1))^{1/2}$.
(9) Let $\widehat{U}^{(s)}$ be the k right singular vectors of $\widehat{E}^{(s\ell)}$.
(10) Set $\widehat{\beta} = \widehat{\Sigma}\widehat{\beta}^{(s)}$, $\widehat{U} = \widehat{\Sigma}\widehat{U}^{(s)}$.
(11) Set $Y_i^{(p)} = (Y_{i1}^{(r)} - \widehat{\beta}_i^\mathsf{T}X_1^{(r)})/\widehat{\sigma}_i$, $U_i^{(p)} = \widehat{U}_i^{(s)}$.
(12) Fit Θ–IPOD with response $Y_i^{(p)}$ predictors $U_i^{(p)}$ getting $\widehat{\gamma}_i^{(p)}$, $\widehat{V}_1^{(p)}$ and $\widehat{\tau}$.
(13) Let $T_i = (Y_i^{(p)} - U_i^{(p)\mathsf{T}}\widehat{V}_1^{(p)})/\widehat{\tau}$, $i = 1, \ldots, N$.
(14) Rank genes from most significant (largest $|T_i|$) to least.

FIG. 1. *The LEAPP algorithm, using notation from the text. Step (6) can be omitted if the desired value of k is already known. Step (8)(d) is written concisely but can be computed more efficiently. We use $|T_i|$ to rank genes. Convergence at (8) is declared when $\|\widehat{\Sigma}_{\mathrm{new}} - \widehat{\Sigma}_{\mathrm{old}}\|_1/\|\widehat{\Sigma}_{\mathrm{old}}\|_1 < 10^{-4}$ with $\|\cdot\|_1$ here being the sum of absolute diagonal elements. There is an R package for LEAPP at* http://cran.r-project.org/web/packages/leapp/.

row of $Y^{(\ell)}$ by taking $\tau^2 = \mathbb{E}((n-1-s)/\chi^2_{n-1-s}) = (n-s-1)/(n-s-3)$. A further approximate adjustment for estimating k latent vectors is to take $\tau^2 = (n-s-k-1)/(n-s-k-3)$. This estimate of τ can be used in (2.13) for ranking of hypotheses if γ is not suspected to be sparse.

2.5. *SVA.* We compare our method to the surrogate variable analysis (SVA) method of Leek and Storey (2008). Their iteratively reweighted surrogate variable analysis algorithm adjusts for latent variables before doing a regression. But it does not isolate them.

A full and precise description of SVA appears in the supplementary information and online software for Leek and Storey (2008). Here we present

a brief outline. Their model takes the form

$$Y = \gamma g^{\mathsf{T}} + UV^{\mathsf{T}} + \Sigma E,$$

where UV^{T} is their "dependence kernel" and E is not necessarily normally distributed but has independent rows.

The SVA algorithm uses iteratively reweighted SVDs to estimate U, V and γ. The weights are empirical Bayes estimates of $\Pr(\gamma_i = 0, U_i \neq 0 \mid Y, g, V)$ from Storey, Akey and Kruglyak (2005). Their method seeks to remove the primary term γg^{T} by downweighting rows with $\gamma_i \neq 0$. Our method creates columns that are free of the primary variable by rotation.

The SVA iteration is as follows. First, they fit a linear model without any latent variables, getting estimates $\widehat{\gamma}$ and the residual $R = Y - \widehat{\gamma} g^{\mathsf{T}}$. Second, they apply the simulation method of Buja and Eyuboglu (1992) to R to estimate the number k of factors, and then take the top k right eigenvectors of R as the initial estimator $\widehat{V}$. Third, they form the empirical Bayes estimates $w_i = \Pr(\gamma_i = 0, U_i \neq 0 \mid Y, g, \widehat{V})$ from Storey, Akey and Kruglyak (2005). Fourth, based on those weights, they perform a weighted singular value decomposition of the original data matrix Y, where row i is weighted by w_i. The weighted SVD gives them an updated estimator $\widehat{V}$. They repeat steps (3) and (4), revising the weights w_i and then the matrix $\widehat{V}$, until $\widehat{V}$ converges. They perform significance analysis on γ through the multivariate linear regression model

$$Y = \gamma g^{\mathsf{T}} + U\widehat{V}^{\mathsf{T}} + \Sigma E,$$

where $\widehat{V}$ is treated as known covariates to adjust for the primary effect g.

To estimate the number k of factors in the SVD, they use a simulation method of Buja and Eyuboglu (1992). That algorithm uses Monte Carlo sampling to adjust for the well-known problem that the largest singular value in a sample covariance matrix is positively biased. That method has two parameters: the number of simulations employed and a significance threshold. The default significance threshold was 0.1 and the default uses 20 permutations.

2.6. *EIGENSTRAT.* EIGENSTRAT [Price et al. (2006)] was developed to control for differences in ancestry in genetic association studies, where the matrix Y represent the alleles carried by the subjects at the genetic markers (e.g., $Y_{ij} \in \{0, 1, 2\}$ counts the number of one of the alleles). The primary variable can be case versus control, disease status or other clinical traits.

In our notation, they begin with a principal components analysis approximating Y by $\widehat{U}\widehat{V}^{\mathsf{T}}$ for $\widehat{U} \in \mathbb{R}^{N\times k}$ and $\widehat{V} \in \mathbb{R}^{n\times k}$. Then for $i = 1, \ldots, N$ they test whether $Y_{i,1:n}$ is significantly related to g in a regression including the k columns of $\widehat{V}$ or, equivalently, whether the partial correlation of $Y_{i,1:n}$ on g, adjusted for $\widehat{V}$, is significant. Although the data are discrete and the

method resembles one for Gaussian data, the results still clearly obtain latent variables showing a natural connection to the geographical region of the subjects' ancestors.

EIGENSTRAT has an apparent weakness. If the signal γg^T is large, then its presence will corrupt the estimates of $\widehat{U}$ and $\widehat{V}$. The estimate $\widehat{V}$ will be correlated with the effect g that we are trying to estimate a coefficient for. Indeed, we find in our simulations of Section 3 that EIGENSTRAT performs poorly when the signal is large compared to the latent variable. While EIGENSTRATs strong latent with weak signal assumption seems to be appropriate for genetic association studies, a method that does not rely on such assumptions is desirable.

EIGENSTRAT also requires the choice of a rank k for the latent term. Price et al. (2006) describe a default choice of $k = 10$. Patterson, Price and Reich (2006) apply a spiked covariance model test of Johnstone (2001) using the Tracy–Widom distribution [Tracy and Widom (1994)].

2.7. *Other methods.* We have used Eigenstrat and SVA in our comparisons because they are widely used in applications. A number of other methods have been proposed for this problem. It is not feasible to include them all in our numerical comparisons. Instead we describe several of them here, relating their approaches to the notation of Section 2.1.

Friguet, Kloareg and Causeur (2009) model their data as $Y = \gamma g^\mathsf{T} + UV^\mathsf{T} + \Sigma E$. They assume the latent V is normally distributed (independent of E) and that U is nonrandom. They do not assume sparsity for γ. They estimate U, V, γ and Σ by an EM algorithm. They find that using $\widehat{V}$ in an FDR procedure is an improvement compared to a model that does not employ latent variables.

Lucas, Kung and Chi (2010) take $Y = \beta X^\mathsf{T} + UV^\mathsf{T} + \Sigma E$ and make extensive use of sparsity priors. They include the primary variable g as one of the columns of X, instead of singling it out as we do. Under their sparsity priors, a coefficient is either 0 or it is $\mathcal{N}(0, \tau^2)$. The probability of a nonzero coefficient is π, which in turn has a Beta distribution with a small mean. They apply sparsity priors to the elements of both the coefficient matrix β and the latent variables U. The parameters π and τ are different for each column of β. They use Markov chain Monte Carlo for their inferences.

Allen and Tibshirani (2010) model the data as $Y = \gamma g^\mathsf{T} + E$ where $E \sim \mathcal{N}(0, \Sigma \otimes \Gamma)$. That is, the noise covariance is of Kronecker form which models dependence between rows and between columns. Our model has a different variance equal to the sum of two Kronecker matrices, one from UV^T and one from ΣE. They estimate their Σ and Γ by maximum likelihood with a penalty on the norm of the inverses of Σ and Γ. Their L_1 penalties encourage sparsity in $\widehat{\Sigma}^{-1}$ and $\widehat{\Gamma}^{-1}$. They then whiten Y using $\widehat{\Gamma}$ and $\widehat{\Sigma}$ and apply false discovery rate methods. They also show that correlations among different

columns lead to incorrect estimates of FDR, while correlated rows do not much affect the estimates of FDR.

Efron (2007) proposed a method to fit an empirical null to the data to directly account for correlations across arrays. The empirical null method works with estimated Z scores (one per gene) and uses the histogram of those scores to account for the effects of latent variables. This process adjusts significance levels for hypotheses but does not alter their ordering.

Carvalho et al. (2008) consider similar problems but apply a very different formulation. They treat the primary variable (our g) as the response and use the data matrix (our Y) as predictors.

2.8. *Rank estimation.* The problem of choosing the number k of latent variables is a difficult one that arises for all the methods we used. The Tracy–Widom strategy is derived for the case with $\Sigma = \sigma I_N$, while our motivating applications have heteroscedasticity.

Even for $\Sigma = \sigma I_N$ it is known that the best rank for estimating UV^T is not necessarily the true rank. There is a well-known threshold strength below which a factor is not detectable and Perry (2009) shows that there is a still higher threshold below which estimating that factor worsens the estimate of UV^T. Owen and Perry (2009) present a cross-validatory estimate for the rank k and Perry (2009) shows how to tune it to choose a rank k that gives the best reconstruction as measured by the Frobenius norm.

In our numerical comparisons, LEAPP, SVA and EIGENSTRAT were all given the same rank k to use. Sometimes k was fixed at a default value. Other times we used the method of Buja and Eyuboglu (1992).

3. Performance on synthetic data. In this section we generate data from the model (2.1) and compare the results from the algorithms to each other, to an oracle which is given the latent variable, and to a raw regression method which makes no attempt to adjust for latent variables. Some simulations by Sun (2011) made under a different model are described in Section 5.

We choose $s = 0$, omitting the βX^T covariate term, so the simulated data satisfy

$$Y = \gamma g^\mathsf{T} + UV^\mathsf{T} + \Sigma E. \tag{3.1}$$

The model (3.1) is a special case of both the LEAPP model and the SVA model.

Our simulations have $n = 60$ (subjects) and $N = 1000$ (genes). Our primary covariate is a binary treatment vector $g \propto (1, \ldots, 1, -1, \ldots, -1)$, with equal numbers of 1 and -1, normalized so that $g^\mathsf{T} g = 1$.

The vector γ of treatment effects has independent components γ_i taking the values $c > 0$ and 0 with probability $\pi = 0.1$ and $1 - \pi = 0.9$, respectively. We chose c in order to attain specific signal to noise ratios as described below. The matrix Σ is a diagonal with nonzero entries σ_i sampled indepen-

dently from an inverse gamma distribution: $1/\sigma_i^2 \sim \text{Gamma}(5)/4$. Note that $\mathbb{E}(\sigma_i^2) = 1$.

We use $k = 1$ latent variable that has correlation ρ with g. The latent vector $U = (u_1, \ldots, u_N)$ is generated as independent $U(-a, a)$ random variables. We will choose a to obtain specific latent to noise variance ratios. The latent vector V is taken to be $\rho g + \sqrt{1-\rho^2}W$, where W is uniformly distributed on the set of unit vectors orthogonal to g. That is, we sample V so as to have a sample correlation and squared norm that both match their population counterparts.

The model (3.1) gives Y three components: the signal $\mathcal{S} = \gamma g^\mathsf{T}$, the latent structure $\mathcal{L} = UV^\mathsf{T}$, and the noise $\mathcal{N} = \Sigma E$. The relative sizes of these components affect the difficulty of the problem. We use Frobenius and spectral norms to describe the sizes of these matrices.

The noise matrix is constructed so that $\mathbb{E}(\sigma_i^2 \varepsilon_{ij}^2) = \mathbb{E}(\sigma_i^2) = 1$, so that $\mathbb{E}(\|\mathcal{N}\|_F^2) = Nn$. Because the signal and latent matrices have rank 1,

$$\mathbb{E}(\|\mathcal{S}\|_F^2) = \mathbb{E}(\|\mathcal{S}\|_2^2) = \mathbb{E}(\|\gamma\|_2^2) = N\pi c^2 \tag{3.2}$$

and

$$\mathbb{E}(\|\mathcal{L}\|_F^2) = \mathbb{E}(\|\mathcal{L}\|_2^2) = \mathbb{E}(\|U\|_2^2) = Na^2/3. \tag{3.3}$$

For our simulation, we specified the ratios

$$\text{SNR} \equiv \pi c^2 \quad \text{and} \quad \text{LNR} \equiv a^2/3$$

and varied them over a wide range. We also use $\text{SLR} = 3\pi c^2/a^2$.

We also varied the level of ρ, the correlation between the latent and primary variables. For each setting of SNR, SLR, LNR and ρ under consideration, we simulated the process 100 times and prepared ROC curves, from the pooled collection of 100,000 predictions.

The methods that we applied are as follows:

true	an oracle given UV^T which then does regression of $Y - UV^\mathsf{T}$ on g,
raw	multivariate regression of Y on g ignoring latent variables,
eig	EIGENSTRAT of Price et al. (2006),
sva	surrogate variable analysis from Leek and Storey (2008), and
lea	our proposed LEAPP method.

The ROC curves for two sets of conditions are shown in Figure 2. The best performance is always from the oracle. The next best method is LEAPP. For the conditions in the left panel RAW is next best followed by SVA and EIGENSTRAT. In the right panel SVA is third, followed by EIGENSTRAT and then RAW.

Because the ROC curves from the simulations have few if any crossings, we can reasonably summarize each one by a single number. We have used the area under the curve (AUC) for a global comparison. We also use a precision measure for the quality of the most highly ranked values. That

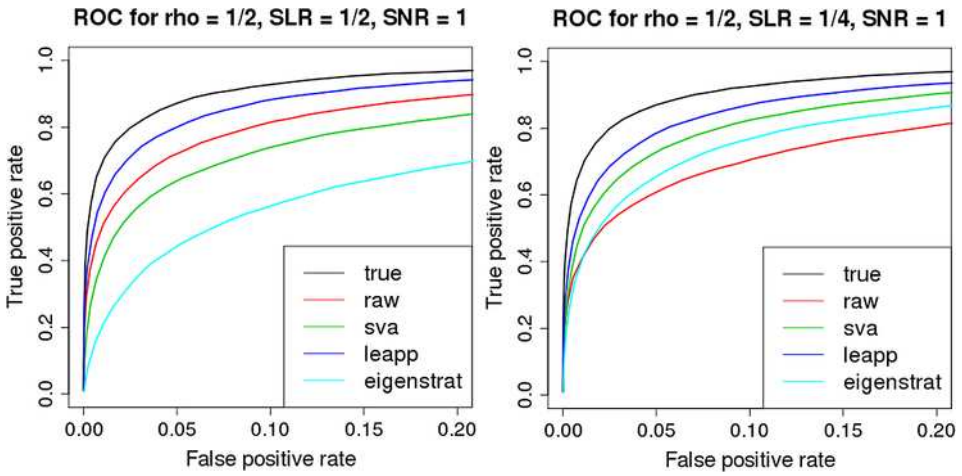

FIG. 2. *This figure shows the knee of the ROC curves for two simulations with* $\rho = 1/2$ *and* $SNR = 1$. *The left panel has* $SLR = 1/2$. *In this case the raw method beats SVA which beats EIGENSTRAT. The right panel has* $SLR = 1/4$ *and SVA beats EIGENSTRAT which beats the raw method. In every case we simulated, the best results are for an oracle that was given the latent variables. The second best was always for the proposed LEAPP method. The relative performance for SVA, EIGENSTRAT and the raw method were different in other settings.*

measure is the fraction of truly nonnull genes among the highest ranking H genes. We use $H = 50$.

When $\rho = 0$, EIGENSTRAT, SVA and LEAPP have almost equivalent performance. For $\rho > 0$, the oracle always had the highest AUC and LEAPP was always second. The ordering among the other three methods varied. Sometimes EIGENSTRAT was the best of those three, other times SVA was the best of those three and other times RAW was the best of those three.

Figure 3 shows a heatmap of the improvement in AUC for LEAPP versus SVA. The improvements are greatest when ρ is large. This is reasonable because SVA is not designed to account for correlation between the latent and primary variables. At each correlation level, the greatest differences arise when SNR is small and LNR is about 2.

Figure 4 shows the improvement in AUC for LEAPP versus EIGENSTRAT. The improvements are largest when the primary effect is large.

The improvements versus SVA are smaller than those versus EIGENSTRAT. To judge the practical significance of the improvement, we repeated some of these simulations for SVA, increasing n until SVA achieved the same AUC that LEAPP did. Sometimes SVA required only 2 more observations (one treatment and one control) to match the AUC of LEAPP. Sometimes it was unable to match the AUC even given double the sample size, that is, $n = 120$ observations instead of $n = 60$. Not surprisingly, the advantage of

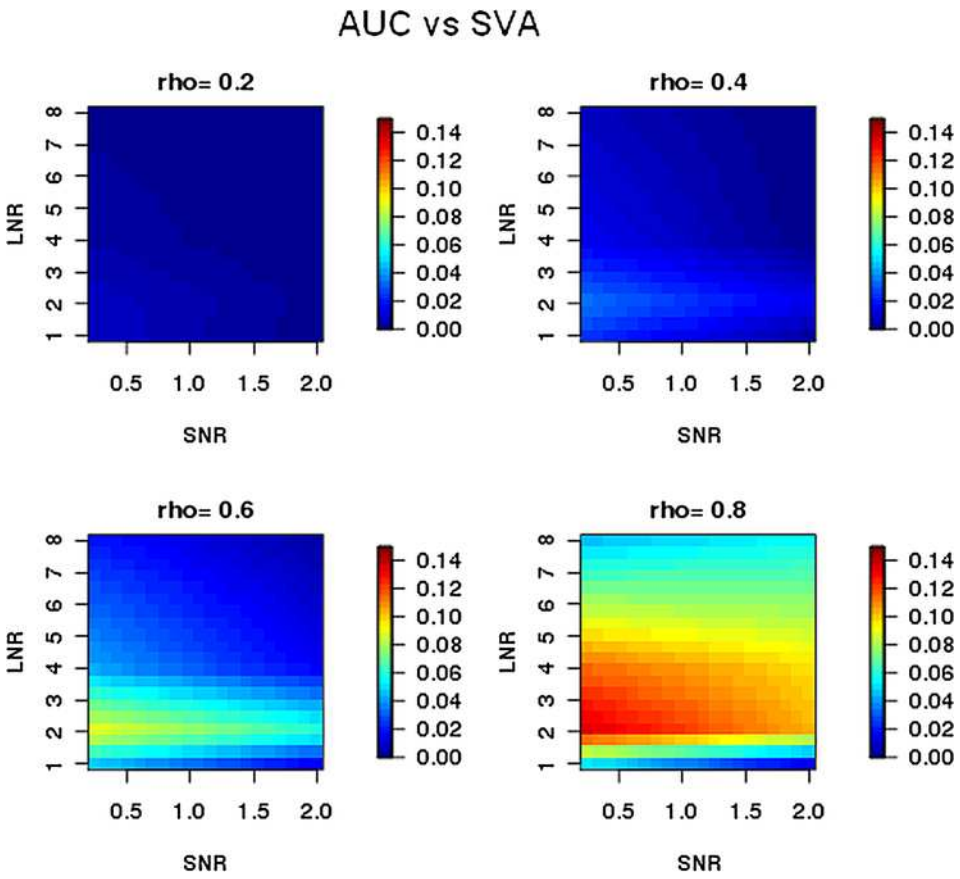

FIG. 3. *This figure shows the improvement in AUC for LEAPP relative to SVA. Here ρ is the correlation between the primary and latent variables. The signal to noise ratio and latent to noise ratio are described in the text. The color scheme encodes* $(\mathrm{AUC}_{\mathrm{lea}} - \mathrm{AUC}_{\mathrm{sva}})/\mathrm{AUC}_{\mathrm{sva}}$.

LEAPP is greatest when the latent variable is most strongly correlated with the primary.

Table 1 shows a feature of this problem that we also see in the figures. The improvement over SVA is quite small when $\mathrm{LNR} = 0.5$. A small enough latent effect becomes undetectable, both methods suffer and there is little difference. Similarly, a very large latent effect ($\mathrm{LNR} = 8$) is easy to detect by both methods. The largest differences arise for medium sized latent effects.

High throughput methods are often used to identify candidates for future follow-up investigation. In that case we value high precision for the most highly ranked hypotheses. Figure 5 shows the improvement of LEAPP over SVA, as measured by precision. Figure 6 shows the improvement of LEAPP over EIGENSTRAT, as measured by precision.

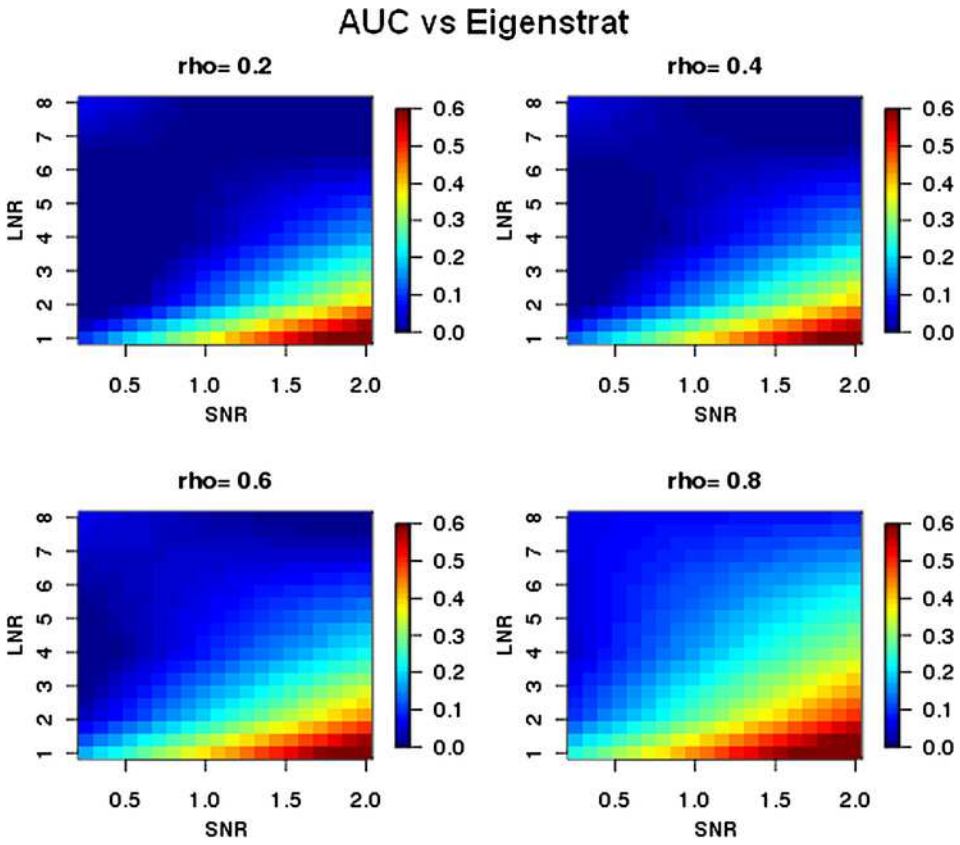

FIG. 4. *This figure shows the improvement in AUC for LEAPP relative to EIGENSTRAT. The simulation conditions are as described in Figure 3. The color scheme encodes* $(\mathrm{AUC}_{\mathrm{rot}} - \mathrm{AUC}_{\mathrm{eig}})/\mathrm{AUC}_{\mathrm{eig}}$.

4. AGEMAP data. It is hard to find a real data set where the true set of important genes is known. Even if we are confident that a few genes are active, we still cannot be sure that the others are really inactive: the corresponding null hypotheses might be accepted, but they are not proved. We turn instead to the AGEMAP study [Zahn et al. (2007)].

The AGEMAP study [Zahn et al. (2007)] investigated age-related gene expression in mice. Ten mice at each of four age groups were investigated. From these 40 mice, samples were taken of 16 different tissues, resulting in 640 microarray data sets. A small number of those 640 microarrays were missing. From each microarray, 8932 probes were sampled. Perry and Owen (2010) found that many of the tissues in this data set exhibited strong latent variables. Their approach assumed that the latent variables were orthogonal to the treatment.

TABLE 1

This table shows the number of samples required for SVA to attain the same AUC that LEAPP attains with $n = 60$ samples. For example, with $SNR = 2$ and $LNR = 0.5$, and $\rho = 0.25$, SVA requires 66 samples or 10% more. The entries of 100% denote settings where the increase needed was $\geq 100\%$

Conditions		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
SNR	**LNR**	n	%	n	%	n	%
2	0.5	66	10	66	10	62	3
2	1	68	13	92	53	120	100
2	2	66	10	74	23	114	90
2	4	62	3	66	10	88	47
2	8	62	3	66	10	72	20
1	0.5	64	7	64	7	62	3
1	1	66	10	90	50	120	100
1	2	64	7	76	27	120	100
1	4	64	7	66	10	90	50
1	8	62	3	66	10	76	27
0.5	0.5	64	7	64	7	62	3
0.5	1	66	10	84	40	120	100
0.5	2	66	10	78	30	110	83
0.5	4	66	10	68	13	88	47
0.5	8	62	3	68	13	72	20

Our underlying assumption is that aging should have partially though not totally consistent results from tissue to tissue. According to Kim (2008): "Some aspects of aging only affect specific tissues; examples include progressive weakness of muscle, declining synaptic function in the brain, and decreased filtration rate in the kidney. Other aspects of aging occur in all cells regardless of their tissue type, such as the accumulation of oxidative damage, and telomere shortening." Zahn et al. (2006) found some genetic pathways with common age regulation in (human) kidney, brain and muscle. Rodwell et al. (2004) found common aging between human kidney, cortex and medulla. Some aspects of aging are also common from species to species Kim (2007).

A tendency for some common component to aging should in turn produce overlap in gene lists computed from multiple tissues. Because age-related genes are sparse, noisy estimation is more likely to reduce overlap in gene lists than to create it.

To illustrate this point, consider a setting with 1000 genes and two tissues A and B with counts

$$\begin{array}{c} \\ B \\ \neg B \end{array} \begin{array}{c} \begin{array}{cc} A & \neg A \end{array} \\ \begin{pmatrix} 10 & 10 \\ 10 & 970 \end{pmatrix} \end{array}.$$

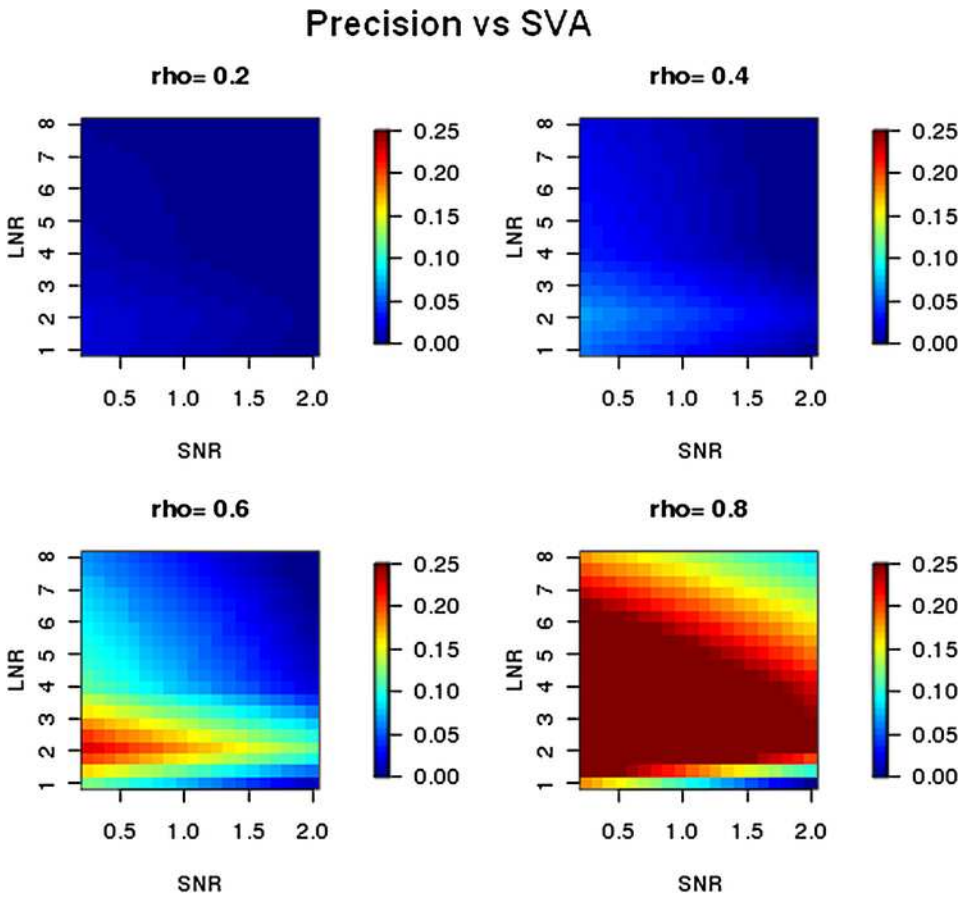

FIG. 5. *This figure shows the improvement in precision for LEAPP relative to SVA. Precision is the fraction of truly affected genes among the top* $H = 50$ *ranked genes. The simulation conditions are as described in Figure 3. The color scheme encodes* $(\mathrm{PRE}_{\mathrm{lea}} - \mathrm{PRE}_{\mathrm{sva}})/\mathrm{PRE}_{\mathrm{sva}}$.

Here 10 genes are truly age-related in both tissues, 10 are age-related in A but not B, and, finally, 970 genes are not age-related in either tissue. Suppose now that statistical testing identifies each truly age-related gene with power 0.6 and that each nonage-related gene has a false discovery probability of 0.01. Using $\widehat{A}$ and $\widehat{B}$ to represent genes identified as age-related, the expected counts (for independent test statistics) are in the following matrix:

$$\begin{array}{c} \\ \widehat{B} \\ \neg\widehat{B} \end{array} \begin{array}{c} \begin{array}{cc} \widehat{A} & \neg\widehat{A} \end{array} \\ \begin{pmatrix} 3.817 & 17.983 \\ 17.983 & 960.217 \end{pmatrix} \end{array}.$$

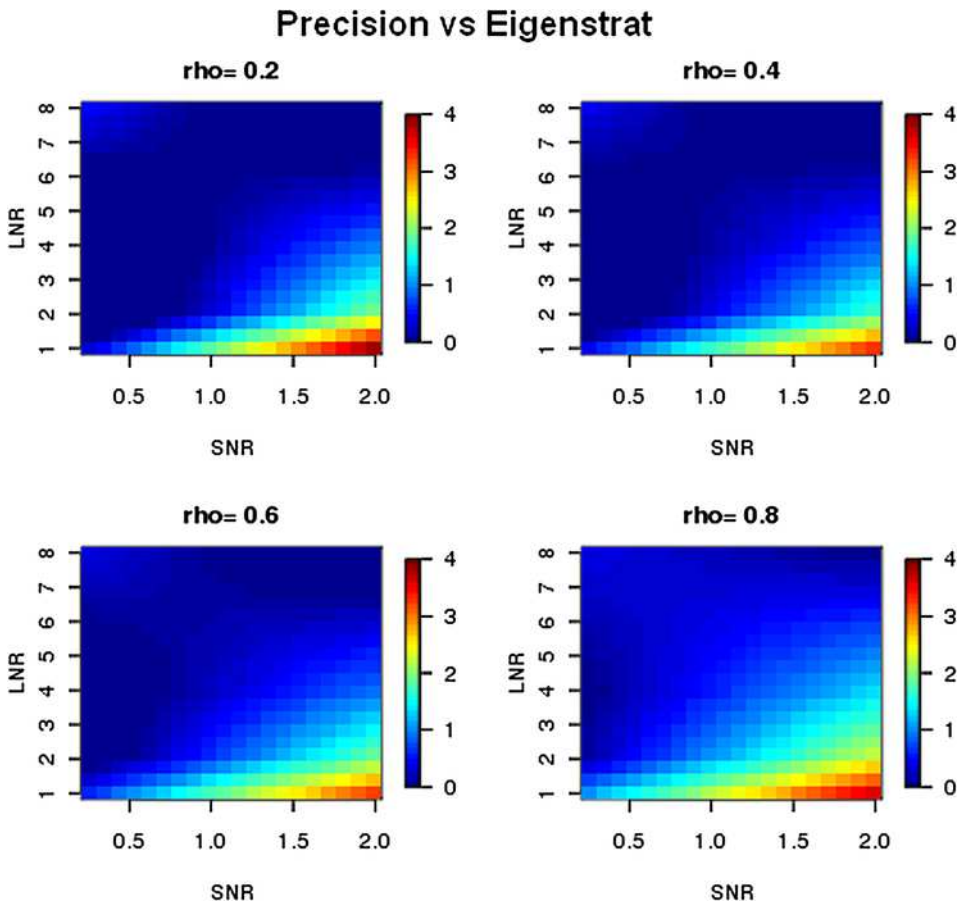

FIG. 6. *This figure shows the improvement in precision for LEAPP relative to EIGENSTRAT. Precision is the fraction of truly affected genes among the top* $H = 50$ *ranked genes. The simulation conditions are as described in Figure 3. The color scheme encodes* $(\mathrm{PRE}_{\mathrm{rot}} - \mathrm{PRE}_{\mathrm{eig}})/\mathrm{PRE}_{\mathrm{eig}}$.

The effect of noisy gene identification is severely biased toward reducing the apparent overlap.

For any two tissues, we can measure the overlap between their sets of highly ranked genes. For two sets A and B, their resemblance [Broder (1997)] is

$$\mathrm{res}(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where $|\cdot|$ denotes cardinality. Given two tissues and a significance level α, we can compute the resemblance of the genes identified as age-related in the tissues. Resemblance is then a function of α. Plotting the numerator $|A \cap B|$ versus the denominator $|A \cup B|$ as α increases, we obtain curves depicting the strength of the overlap.

In our setting with 16 tissues there are $\binom{16}{2} = 120$ resemblances to consider. To keep the comparison manageable as well as to pool information from all tissues, we computed the following quantities:

$$I_\alpha = \sum_{1 \le j < j' \le 16} |A_j^\alpha \cap A_{j'}^\alpha| \quad \text{and} \quad U_\alpha = \left| \bigcup_{j=1}^{16} A_j^\alpha \right|, \tag{4.1}$$

where A_j^α is the set of statistically significant genes at level α for tissue j. We can think of I_α/U_α as a pooled resemblance. We would like to see large I_α at each given level of U_α.

Figure 7 plots I_α versus U_α for the methods we are comparing. To make a precise comparison, we arranged for each method that estimated latent structure to employ the same estimate for the rank of the latent component. That rank is either 1, 2, 3 or the value chosen by the method of Buja and Eyuboglu (1992). At any rank LEAPP generates the most self-consistent gene lists over almost the entire range. EIGENSTRAT is usually second. SVA beats a raw method that makes no adjustments. LEAPP retains its strong performance when the rank is chosen from the data while the other two methods become poorer in that case.

Resemblance across tissues could also be high if there exists latent variables strongly correlated with age which are repeated across tissues. For example, consider a scenario where all tissues from young mice are in one batch, and all tissues from elder mice are in a different batch. If there are strong batch biases, then "age-related" genes would be reported by the raw method, and the same genes would be ranked high across all tissues. However, note that raw performs the worst of all methods in Figure 7, which gives some reassurance that the high resemblance of the other methods is due to successful removal of latent variables.

Given what we have learned from simulations, the relative performance of EIGENSTRAT and SVA gives us some insight into these data. Since EIGENSTRAT has done well, it is more likely that the signal is not very strong. Since SVA has done poorly, it is more likely that the latent variables in these data are correlated with age. There is also the possibility that they are correlated with sex (the covariate). Our simulations did not include a covariate.

5. Conclusions. High throughput testing has performance that deteriorates in the presence of latent variables. Latent variables that are correlated with the treatment variable of interest can severely alter the ordering of p-values. Our LEAPP method separates the latent variable from the treatment variable, making an adjustment possible.

We have found in simulations that the adjustment brings about a better ordering among hypotheses than is available from either SVA or EIGENSTRAT. The improvement over SVA is largest when the latent variable is

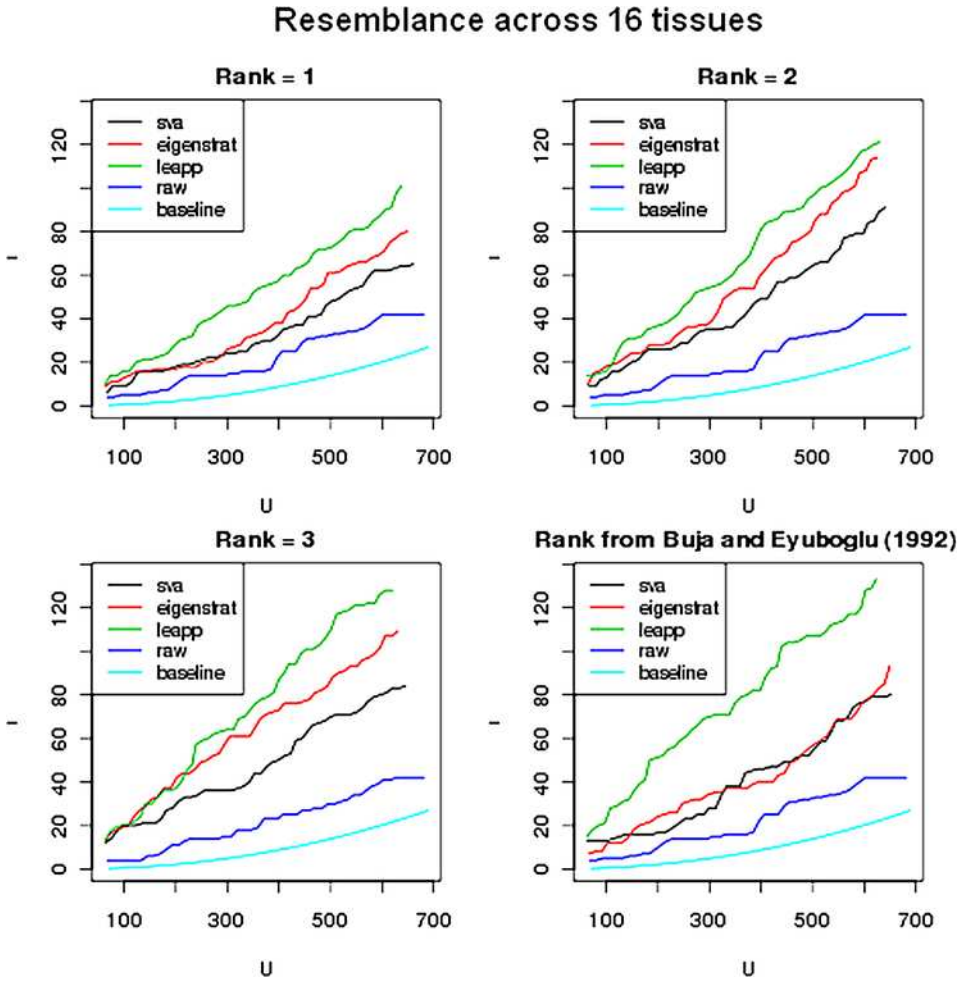

FIG. 7. *This figure shows the resemblance among significant gene sets from* 16 *tissues in the AGEMAP study. We plot* I_α *versus* U_α *[from equation (4.1)], increasing* α *from* 0 *until* $U_\alpha = 700$. *The greatest self-consistency among lists is from LEAPP. EIGENSTRAT is second best. The baseline curve is computed assuming that the rankings for all* 16 *tissues are mutually independent.*

correlated with the primary one. The improvement over EIGENSTRAT is largest when the primary variable has a large effect.

A referee asked about the case where the coefficients of γ for the primary variable correlate over genes with the per gene latent variable, U in our notation. We have not simulated such a case. It might be very difficult for all methods or it might be comparable to the case where g correlates with V. It seems clear that if UV^T matches γg^T closely enough, then it will be impossible to identify relevant genes in this model.

In the simulations reported here the data are drawn from the model under which LEAPP was derived. Sun (2011) also simulates the LEAPP, SVA and EIGENSTRAT algorithms on the model used by Price et al. (2006) to represent SNP association studies. The SNPs themselves are drawn from the Balding–Nichols model [Balding and Nicols (1995)]. Two scenarios were considered. In both, the LEAPP ROC curve placed above that for SVA which was above that for EIGENSTRAT. All methods were close when the relative risk for the causal allele was $R = 1.5$ while EIGENSTRAT lagged behind for the case with $R = 3$.

On the AGEMAP data we found better consistency among tissues for significance estimated by LEAPP than for either SVA or EIGENSTRAT.

Some applications may have features measured on the genes with per-sample covariates to be estimated statistically. Such terms can be included in the criss-cross regression framework but we have no experience fitting them.

LEAPP produces p-values in addition to the relative ordering of the genes. In this paper we have only looked at the quality of the relative ordering. In response to a reviewer's query about calibration of p-values, we created a QQ-plot of test statistics T_i at (2.13) on simulated data (not shown) and found it very nearly linear. That simulated data was pure noise, having no regression or latent structure. For an investigation on real data, Sun [(2011), Chapter 4.5.2] considered the breast cancer data from Hedenfalk (2001). She finds that the test statistics produced by LEAPP have an empirical null distribution from the R package locfdr [Efron (2008)] of $\mathcal{N}(0.012, 1.018^2)$ that closely matches the nominal null distribution. That is what we would expect to see if the nominal p-values coming out of LEAPP had the $U[0,1]$ distribution that they should have. Corresponding empirical nulls are $\mathcal{N}(-0.01, 1.55^2)$ for the RAW method, $\mathcal{N}(-0.009, 1.425^2)$ for SVA and $\mathcal{N}(-0.093, 1.199^2)$ for EIGENSTRAT. Thus, in addition to a general improved ordering of genes, this one example had p-values that are better calibrated in LEAPP than in SVA or EIGENSTRAT.

APPENDIX

Here we give some properties of our approach to testing many hypotheses in the presence of latent variables. We focus on a simpler version of the model that is more tractable:

$$Y = \gamma g^{\mathsf{T}} + UV^{\mathsf{T}} + \sigma E, \tag{A.1}$$

where $g \in \mathbb{R}^{n\times 1}$ with $\|g\| = 1$ as before, $U \in \mathbb{R}^{N\times k}$ is nonrandom, $V \in \mathbb{R}^{n\times k}$ has IID rows with $\mathbb{E}(V^{\mathsf{T}}V) = I_k$, known rank k and $E \sim \mathcal{N}(0, I_N \otimes I_n)$. Compared to the full model (2.1), equation (A.1) has no covariate term βX^{T}, and has constant variance $\Sigma = \sigma I_N$.

This simplification allows us to apply results from the literature to our model. It removes the Monte Carlo based rank estimation step and the alternation between estimating Σ and using the estimate $\widehat{\Sigma}$. When $k=1$, the primary to latent correlation is $\rho = g^{\mathsf{T}}V/\sqrt{V^{\mathsf{T}}V}$.

Our algorithm requires the choice of a rotation matrix O such that $Og = e_1$. There are multiple possibilities for this matrix. Our algorithm is invariant to the choice of O.

THEOREM A.1. *Let Y follow the model (A.1). Then our estimates of U and γ do not depend on the rotation O used as long as $Og = e_1$.*

PROOF. See Sun (2011). □

It is not hard to extend the proof of Theorem A.1 to account for the βX^{T} term. The criss-cross regression begins by computing $\widehat{\beta}$ from sums of squares and cross-products. Those sums of squares and cross-products are invariant under the rotation.

The following theorem provides a sufficient condition for our estimate $\widehat{U}$ to consistently estimate U. We study the case where the data are generated with $k=1$ and the model is also estimated using the correct rank $k=1$. Then as long as the latent factor U is large enough compared to the noise level, we will be able to detect and estimate U fairly well. Our size measure $\|U\|_2^2(1-\rho^2)/n$ takes account of the correlation. With a higher ρ, more of the latent factor is removed from $Y^{(\ell)}$.

We measure error by the cosine $\Phi(\widehat{U}, U) = \widehat{U}^{\mathsf{T}}U/(\|\widehat{U}\|_2\|U\|_2)$ of the angle between $\widehat{U}$ and U. The estimate $\widehat{U}$ is determined only up to sign. Replacing $\widehat{U}$ by $-\widehat{U}$ causes a change from $\widehat{V}$ to $-\widehat{V}$ and leaves the model unchanged. We only need $\max(\Phi(\widehat{U}, U), \Phi(-\widehat{U}, U)) = |\Phi(\widehat{U}, U)| \to 1$ for consistency.

THEOREM A.2. *Let Y follow the model (A.1) with $k=1$ and $\|U\|_2^2(1-\rho^2)/n \to \infty$ and $N(n)/n \to c \in (0,\infty)$ as $n \to \infty$. Let $\widehat{U}$ be our estimator for U using $k=1$. Then $|\Phi(\widehat{U}, U)| \to 1$ as $n \to \infty$ with probability 1.*

PROOF. See Sun (2011). □

Next we give conditions for the final step of LEAPP to accurately estimate γ, that is, for $\|\widehat{\gamma} - \gamma\|_2$ to be small. To do this, we combine methods used in random matrix theory from Bai (2003) with methods used in compressed sensing in Candès and Randall (2006).

In our simulations we found little difference between robust and nonrobust versions of the Θ–IPOD algorithm. This is not surprising, since our simulations did not place nonzero γ_i preferentially at high leverage points (extreme U_{i1}). For our analysis we replace the robust Θ–IPOD algorithm by the Dantzig selector for which strong results are available.

Our algorithm was designed assuming that the primary variable g is not too strongly correlated with the latent variable V. In our analysis we also impose a separation between the effects γ and the latent quantity U. Specifically, we assume that γ is sparse and that U is not.

The vector x is s-sparse if it has at most s nonzero components. Following Candès and Randall (2006), we define the sequences $a_s(A)$ and $b_s(A)$ as the largest and smallest numbers, respectively, such that

$$a_s(A)\|x\|_2 \le \|Ax\|_2 \le b_s(A)\|x\|_2$$

holds for all s-sparse x.

THEOREM A.3. *Suppose that Y follows the model (A.1) with $k = 1$, a fixed correlation $\rho \in (-1, 1)$ between g and V, and an s-sparse vector γ. Assume that $N/n \to c \in (0, \infty)$, $V^{\mathsf{T}}V \xrightarrow{p} 1$, and $(Nn)^{-1}\|U\|_2^2 \to \sigma_u^2 > 0$ hold as $n \to \infty$. Let our estimated U be $\widehat{U}$ and set $U^\star = \widehat{U}/\|\widehat{U}\|_2$. Writing $|U^\star_{(1)}| \ge |U^\star_{(2)}| \ge \cdots \ge |U^\star_{(N)}|$ for the ordered components of $U^\star$, assume that there is a constant $0 < B < 1$ such that*

$$\sum_{i=1}^{2s} (U^\star_{(i)})^2 + \frac{1}{2}\sum_{i=1}^{3s} (U^\star_{(i)})^2 \le B.$$

Then the Dantzig estimator $\widehat{\gamma}$, which minimizes

$$\|\widehat{\gamma}\|_1 \quad \text{subject to} \quad \|(I - U^\star U^{\star\mathsf{T}})(Y_1^{(r)} - \widehat{\gamma})\|_\infty \le \sigma\sqrt{\log N}$$

satisfies

$$\|\widehat{\gamma} - \gamma\|_2^2 \le \frac{16\sigma^2 s \log(N)}{(1-\rho^2)(1-B)^2}.$$

PROOF. See Sun (2011). □

REFERENCES

ALLEN, G. I. and TIBSHIRANI, R. J. (2010). Inference with transposable data: Modeling the effects of row and column correlations. Technical report, Stanford Univ., Dept. Statistics.

BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857

BALDING, D. and NICOLS, R. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96** 3–12.

BRODER, A. Z. (1997). On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings* 21–29. IEEE Comput. Soc., Los Alamitos.

BUJA, A. and EYUBOGLU, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research* **27** 509–540.

CANDÈS, E. J. and RANDALL, P. A. (2006). Highly robust error correction by convex programming. *IEEE Trans. Inform. Theory* **54** 2829–2840. MR2450835

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. MR2655722

Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion. *Biometrika* **94** 759–771.

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M. and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36** e126.

Dudoit, S. and van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genetics*. Springer, New York.

Efron, B. (2007). Size, power and false discovery rates. *Ann. Statist.* **35** 1351–1377. MR2351089

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

Friguet, C., Kloareg, M. and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571

Gabriel, K. R. and Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* **21** 489–498.

Hedenfalk, I. (2001). Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344** 539–548.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

Kim, S. K. (2007). Common aging pathways in worms, flies, mice and humans. *J. Exp. Biol.* **210** 1607–1612.

Kim, S. K. (2008). Genome-wide views of aging gene networks. In *Molecular Biology of Aging* 215–235. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

Leek, J. T., Scharpf, R. B., Corrada-Bravo, H., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerley, K. and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11** 733–739.

Lucas, J. E., Kung, H. N. and Chi, J. T. A. (2010). Latent factor analysis to discover pathway-associated putative segmental aneuploidies in human cancers. *PLoS Comput. Biol.* **6** e100920:1–15.

Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.

Owen, A. B. and Perry, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat.* **3** 564–594.

Patterson, N. J., Price, A. L. and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* **2** 2074–2093.

Perry, P. O. (2009). Cross-validation for unsupervised learning. Ph.D. thesis, Stanford Univ.

Perry, P. O. and Owen, A. B. (2010). A rotation test to verify latent structure. *J. Mach. Learn. Res.* **11** 603–624. MR2600622

Price, A. L., Patterson, N. J., Plengt, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principal components ananysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38** 904–909.

Rodwell, G., Sonu, R., Zahn, J. M., Lund, J., Wilhelmy, J., Wang, L., Xiao, W., Mindrinos, M., Crane, E., Segal, E., Myers, B., Davis, R., Higgins, J., Owen, A. B. and Kim, S. K. (2004). A transcriptional profile of aging in the human kidney. *PLoS Biology* **2** 2191–2201.

She, Y. and Owen, A. B. (2011). Outlier identification using nonconvex penalized regression. *J. Amer. Statist. Assoc.* **106** 626–639.

Storey, J. D., Akey, J. M. and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology* **3** 1380–1390.

Sun, Y. (2011). On latent systemic effects in multiple hypotheses. Ph.D. thesis, Stanford Univ.

Tracy, C. A. and Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Comm. Math. Phys.* **159** 151–174. MR1257246

Zahn, J. M., Sonu, R., Vogel, H., Crane, E., Mazan-Mamczarz, K., Rabkin, R., Davis, R. W., Becker, K. G., Owen, A. B. and Kim, S. K. (2006). Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS Genetics* **2** 1058–1069.

Zahn, J. M., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A., Weeratna, A. T., Taub, D. D., Gorospe, M., Mazan-Mamczarz, K., Lakatta, E. G., Boheler, K. R., Xu, X., Mattson, M. P., Falco, G., Ko, M. S. H., Schlessinger, D., Firman, J., Kummerfeld, S. K., III, W. H. W., Zonderman, A. B., Kim, S. K. and Becker, K. G. (2007). AGEMAP: A gene expression database for aging in mice. *PLoS Genetics* **3** 2326–2337.

Department of Statistics
Stanford University Sequoia Hall
Stanford, California 94305
USA
E-mail: yunting@stanford.edu
nzhang@stanford.edu
owen@stanford.edu